DOCUMENT RESUME

ED 328 594 TM 016 114

AUTHOR Gregory, Kemp
TITLE More than a Decade's Highlight? The Holistic Scoring Consensus and the Need for Change.
PUB DATE Jan 91
NOTE 29p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 24-26, 1991).
PUB TYPE Viewpoints (120) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Child Development; Cost Effectiveness; Elementary Secondary Education; *Holistic Evaluation; Interrater Reliability; *Research Methodology; *Scoring; Student Evaluation; Testing Problems; *Writing Evaluation

ABSTRACT

A balanced appraisal of holistic scoring of writing is presented via: examination of the present popularity of holistic scoring; analysis of several weaknesses associated with the holistic scoring method; and recommendations for remedying these weaknesses. Six reasons for the popularity of holistic scoring are: (1) relative lack of expense; (2) efficiency; (3) reliability; (4) the appeal of a holistic approach; (5) awareness of child development; and (6) invariance of score point descriptions. The weaknesses of holistic scoring are largely those of: lack of time for test takers or scorers; problems of interrater reliability and score scale consistency; inappropriate developmental principles; and the need for more open communication about the tests and the scoring. The overall thrust of this paper is that the popularity of holistic scoring is justified. Furthermore if its weaknesses are dealt with in a joint fashion by publishers, researchers, and educators, holistic scoring will become a better assessment tool for future research. A 26-item list of references is included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED328594

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

KEMP D. GREGORY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

More than a Decade's Highlight? The Holistic Scoring Consensus and the Need for Change

Kemp Gregory

The Psychological Corporation

Paper Presented at the Annual Meeting
of the
Southwest Educational Research Association

January 24 - 26, 1991

San Antonio, Texas

TM016114

More Than a Decade's Highlight? The Holistic Scoring Consensus and the Need for Change

Ever since Socrates, that most paradigmatic of researchers, the essence of intellectual integrity has been held to be the subjection of cherished beliefs and opinions to the rigors of reason. An important corollary of this rigor has been that seekers of truth should expect and, if possible, welcome findings which raise serious questions about - or even contradict - what they are convinced is the case. In that way, it has been asserted, their vision of reality becomes increasingly enlarged, more in accordance with the facts - and therefore more truthful (Kaufmann, 1958).

Surely, however, since human fallibility is a salient feature of this imperfect world, these searchers (and researchers) can be forgiven the occasional lapse of enjoying the confirmation of a conviction. Sometimes, in fact, such a correspondence between rigorous experiment and strong conviction can indicate that truth is, indeed, being revealed.

Early this fall, while reading a recent article on writing assessment by Professor Brian Huot, I felt that a particular truth - the truth of holistic scoring - was being revealed. In "The Literature of Direct Writing Assessment," Huot argues that in the last decade or so, holistic scoring has proven to be the "most popular" of all direct writing procedures (Huot, 1990). Huot's analysis of research data was a final confirmation of what I had for several years believed to be the state of affairs: that holistic scoring is the most popular assessment tool and that there are some very good reasons for this popularity - reasons which, I hope, will continue to make it the instrument of choice (Breland and Jones, 1984; Faigley, Cherry, Joliffe, & Skinner, 1985; Huot, 1990).

These convictions about holistic scoring have a deep experiential base. From 1985-1989, as either a reader, team leader, or scoring director for Measurement Incorporated, a leading test firm in North Carolina, I was involved in the scoring of 50,000-75,000 papers (I lost the exact count some time ago). During that period, I used approximately three dozen different scoring guides (or "rubrics") to evaluate papers for both state-mandated writing assessment programs and for projects involving major corporations that were employing my firm as a testing subcontractor. The vast majority of the papers were scored holistically. In only one of the projects (TECAT) was analytic scoring the only instrument of evaluation. In the other projects using analytics, the holistic method was always the primary mode of evaluation. In a few projects, a sample of holistically scored papers was scored analytically. Primary Trait scoring was never used.

At first I was struck by the diversity of expectations reflected in the various holistic rubrics and anchor papers that I encountered. However, several years and a number of projects later, the similarities began to seem much more important. I saw that states as diverse as Colorado, New Jersey, Texas, and Connecticut (to name only a few) were in basic agreement that some form of holistic scoring (sometimes called focused, modified, registered, or balanced) was the preferable assessment instrument. When I left the North Carolina firm two years ago to join The Psychological Corporation, I was convinced that just about everyone was scoring papers holistically - and that this popularity was justified.

During the last two years, these convictions have only intensified. Given primary responsibility for the creation of the four interpretive manuals for the Stanford Writing Assessment Program, I researched the writing assessment scene as well as deadlines would allow. I examined publications by state

departments of education and rival publishers and read a number of scholarly articles. Months of study reinforced my feelings about both the popularity and validity of holistic scoring.

I was particularly impressed by two studies that indicated the universal scope of holistic scoring. Most of the Measurement Incorporated projects involved elementary and/or secondary school children. Only one or two were designed to evaluate adult writing holistically. But in Teaching and Assessing Writing, White reported that holistic scoring was quite popular on the collegiate level as well. According to White, a 1982 survey of college English departments showed that almost 90 percent of responding departments were using holistic scoring (White, 1985).

A second, more recent, study was also indicative of universality. In "Decentralized Large Scale Essay Scoring," Auchter and Patience, who are associated with the GED high school equivalency program, wrote about both the addition of an essay to the 1988 GED exam and the instrument chosen to evaluate the essay.

> The 1988 Tests of General Educational Development (GED Tests) include a direct measure of writing as part of the Writing Skills Test. Much of the GED Testing Service's research on the feasibility of including an essay focused on defining and maintaining a holistic score scale and on identifying methods of ensuring score stability and reading reliability across scoring sessions and decentralized sites. (Auchter and Patience, Abstract, 1989)

These studies were a clear indication that holistic scoring was not limited to the writing of younger students. It was, in fact, being applied on a regular basis to the written work of older populations.

Huot's 1990 essay was for me a final confirmation of the popularity of holistic scoring. However, his essay was more than just a positive confirmation. Trying to do justice to dissenting voices, it included

discussion of some negative aspects of holistic scoring - aspects I had been aware of but had, for the sake of the projects under consideration, tried to ignore (Hake, 1986; Stock and Robinson, 1987). This fall, as I began to reflect on the strengths and weaknesses of holistic scoring, it became clear that an updated balanced appraisal was called for, precisely because this instrument of evaluation is so valuable. If its weaknesses are ignored, holistic scoring may lose considerable ground in - or even disappear altogether from - the evaluation arenas of the 90's.

It is the purpose of this paper to give such an updated balanced account. It will do so by 1) examining the present popularity of holistic scoring, a popularity based on its considerable strengths; 2) analyzing several of its distressing weaknesses; and 3) suggesting some far-from-conclusive ways to remedy these weaknesses. The overall thrust of the paper is that the popularity of holistic scoring is justified and that if its weaknesses are dealt with in joint fashion by publishers, researchers, and educators alike, holistic scoring will become an even better assessment tool. It will, in other words, be more than a decade's highlight.

As we begin to examine the popularity of holistic scoring, the first reason that comes to mind is its relatively low costs, especially when these are compared with the developmental expenses involved in the creation of standardized multiple choice tests of linguistic ability (White, 1985). Such tests require extensive blueprint and curriculum-match research. If an outside author is involved, his fees (or hers) have to be included in the budget. Of course, someone must write the items that flesh out the skeletal framework of a test, and that usually entails hiring a cadre of item writers. During the whole period of test development, content editors must review, and often make changes in, the item writers' work. Depending on the scope of the test (the number of grade levels and subtests), the process can take many

months. Finally, the staggering costs of production - including the production of future forms - have to be considered.

Compare all of this with a writing test that is scored holistically. Prompts must be generated, but they are not so many or so involved that a great deal of expense is incurred. The bulk of the cost, at least in the large-scale assessment of writing, has to do with overhead and reader salaries. Overhead costs can be a constant, but reader (or "rater') salaries are not economically crippling because most projects are brief. In rare cases, several hundred readers may be needed to evaluate a large number of essays, but as the annual Texas assessment of writing has shown, hundreds of thousands of papers can be scored in a 6 or 7 week period. After a project is completed, readers, who almost always sign contracts for one project at a time, can be sent home.

Naturally, if readers were paid handsomely, the brevity of projects might not be such a determining factor. However, most readers are temporary employees who are given modest wages and no benefits. Even team leaders, arguably the most important members of the scoring enterprise, receive little more than readers. Scoring directors (or "chief readers") alone are paid fairly well, but when their period of leadership is over, they are often asked to become "mere" readers for other projects - at reader pay. Thus salary costs, while significant, can be held down quite a bit.

Clearly, the issue of economics is closely tied to that of efficiency. In fact, no form of free-writing evaluation is as efficient as holistic scoring (Hogan and Mishler, 1981). In terms of training, readers are able to quickly internalize the score point descriptions through a discussion of their concrete application to carefully selected anchor papers. By the afternoon of the first day of training, readers are ready to score several sets of randomly organized papers representing each score point. No later than day two of

training, raters can competently assess sets of essays that will qualify them for "live scoring," that is, for the rating of papers which have not been pre-scored. Readers usually qualify (or agree with the pre-assigned scores 80% of the time) on the initial set of essays (15-20 papers, depending on the project).

On the first day or two of live scoring, reader production is down a bit, but raters quickly adjust to the varied strategies of a wider spectrum of writers and to the need to read papers more rapidly. Very soon, raters are capable of reading a paper every two minutes, which in a 6 or 7 hour workday, amounts to approximately 200 papers per day. In the large-scale assessment of writing, that is the unwritten holistic norm, and readers who are consistently "poor producers" in this regard are retrained - and sometimes eliminated from the pool of readers. In the majority of cases, minimal retraining is required because raters become more experienced in the course of the project. Corresponding drops in anxiety are observed, too. Both of these factors usually contribute to a rise in the agreement rate of the room as a whole. By and large, then, the "scoring machine" that a room of readers becomes works extremely well - that is, quickly. Naturally, such efficiency helps to keep costs down.

The efficiency of test administration should also be taken into account. Some state departments (Maryland, for example) allow their students an entire day to respond to a writing task that is holistically scored. However, most free-writing tests can be administered in a 45-50 minute class period. In the eyes of many teachers and administrators, this is an extremely attractive feature of the test since it does not significantly disrupt the school day.

These twin principles of economy and efficiency have made substantial contributions to the popularity of holistic scoring. In fact, without them,

holistic scoring would probably have been relegated to an ancillary status - a mode of evaluation best suited to certain research purposes.

Even so, one could claim that these "twin principles" are necessary, but not sufficient, conditions for the success of any given instrument of evaluation. What difference would it make to anyone simply to have a relatively inexpensive test that can be easily administered? Other conditions, such as that of test reliability, must also be met. A number of researchers (White, Huot, and Hogan and Mishler, among others) have commented on the notorious early unreliability of essay scoring. Paul Diederich's Measuring Growth in English is often cited to show the staggering potential for reader disagreement in unstructured scoring sessions (Diederich, 1974). However, the key word here is "unstructured." Since Diederich's 1974 study - and to some degree, no doubt, because of it - the large-scale assessment of writing has become a highly structured and quite reliable enterprise. Presently all successful large-scale assessment of holistically scored written expression exhibits the six constraints enumerated in 1985 by White: controlled essay reading, scoring criteria guide, sample papers, checks on the reading in progress, multiple independent scoring, and evaluation and record keeping (White, 1985).

In fact, even in the early 80's, Hogan and Mishler, citing the work of Cooper and Odell (1977), Diederich (1974), and a previous study of their own, declared that "when the proper conditions are met [in holistic scoring], interscorer reliability of .80 or above can be achieved" (Hogan and Mishler, 1981, P. 11). In this NAEP study, entitled "Relationships Among Measures of Writing Skill," Hogan and Mishler added that "Most researchers agree that this level of reliability is possible, despite a widespread notion to the contrary among laypersons" (Hogan and Mishler, 1981, P. 11).

Ten years later, we may feel that such an assessment is a bit too optimistic, but there is no denying that interscorer reliability in the 70's - and occasionally higher - can be obtained on a regular basis in the large-scale holistic assessment of writing. Practical experience in quite a few projects has confirmed the suspicions of research by showing that such numbers are, in fact, possible. For example, in the dozens of projects I worked on at MI, room agreement rates (involving up to a hundred readers) were consistently in the low to mid 70's. These projects involved a number of different states (Colorado, Connecticut, Louisiana, Oklahoma, Maryland, Mississippi, New Jersey, North Carolina, South Carolina, Tennessee, and Texas), 2 major corporations (Riverside Publishing Company and The Psychological Corporation), and the Bahamian Ministry of Education. While some projects were more difficult than others, reader agreement rates were a constant across projects.

One of these projects - the most ambitious of any I worked on - is especially worthy of some extended discussion. During the summer of 1988, I participated as a scoring director in the evaluation of over 60,000 papers for the National Standardization Program of the Second Edition of the Stanford Writing Assessment Program, which was normed concurrently with the Eighth Edition of the Stanford Achievement Test Series (Stanford Writing Assessment Program, Second Edition: National Norms Booklet, 1990). This project was, for a number of reasons, particularly ambitious. First of all, it called for the holistic scoring of almost a hundred prompts in 4 different types of writing - descriptive, narrative, expository, and persuasive - at nine grade levels (grades 3 - 12). Such an attempt seemed to ask for disaster, since prompt and mode variability has for years been a very difficult problem in the

direct assessment of writing. In addition, since scoring guides for each mode were constructed developmentally (that is, with an eye to the ability of students at given levels), readers had to be able to shift gears with a minimal amount of retraining from one level of expectation to another. Naturally, all of this had to be done as quickly as possible.

The results of this extremely ambitious enterprise can only be seen as a major triumph for holistic scoring. As Wiser, Kelleher, and Rolfe reported in their study of this scoring session, "Interrater reliability for trained observers doing holistic scoring was extremely high" (Wiser, Kelleher, and Rolfe, 1989, P. 24). In fact, seeming to defy the difficulties posed by different topics, modes and grade levels, reliabilities were almost always above .75 and were sometimes above .80. The lowest agreement rate (.67) occurred in the persuasive mode at grade 5; the highest reliability (.92) was found in the same mode at grade 9. The National Norms Booklet (1990) of the Stanford Writing Assessment Program contains a wealth of statistics and correlational data which suggest that a Hogan and Mishler breathtaking leap beyond modest success is always a possibility in the trained holistic scoring of writing.

The fourth good reason for the popularity of holistic scoring is also a pragmatic one, but in a different sense. While economy, efficiency, and reliability are all obviously pragmatic in the sense of practical, reason four has a philosophical dimension that has been best represented in our culture by pragmatists such as William James, who insisted on the primacy of concrete lived experience over disembodied abstraction (Wild, 1970). What does this existential emphasis on real life have to do with holistic scoring? I would say: everything. Because as White declared in Teaching and Assessing Writing, it satisfies a deep-seated, but not always explicitly acknowledged,

"need to see things as units, as complete, as wholes" (White, 1985, P. 18). And as Michael Polanyi and others have argued since the late 1950's, this desire to attend to life in a holistic fashion is based on a fundamental epistemological truth: namely, that holistic apprehension is our primary means of knowing. (Polanyi, 1958) In our culture, James, Polanyi, and the Gestalt psychologists have been powerful pragmatic-existential voices crying out against analytic reductionism, which, as numerous commentators have maintained, has defined the intellectual milieu of the twentieth century (Barrett, 1979; Bronowski, 1956; Matson, 1966).

Several fairly recent studies have helped to confirm the truth of holistic epistemology. For example, Marcia Farr's 1984 review of four research projects about speaking and writing emphasizes that at the earliest levels, language is not learned from subset to subset, structure to structure. Instead, it is under normal conditions "learned and used 'all of a piece,' so to speak - i.e., in a holistic fashion (Farr, 1984, p. 127)." Farr cites in particular the work of Hartse, Burke and Woodward (1982). These researchers studied language acquisition in children three to six years old and found them "engaging in the same literacy processes that adults use" (Farr, 1984, p. 129).

A brief look at one of these "process universals" (which are actually significant learning strategies) will help show what is meant. Hartse, Burke and Woodward refer to a process they label "the fine tuning of language by language." For example, young children assimilate the basic elements of story structure (e.g., the temporal sequence of beginning, middle, and end) by listening to stories. Thus, when they begin to write their own stories, they are able to draw upon their experience to produce, in a parallel fashion, examples of written expression which resemble the stories they have heard.. Adults are engaged in a similar process when they are composing. The difference is one of degree, not of kind.

As odd as it may seem, holistic scoring by mature individuals many years removed from the child's earliest attempts to read and write is clearly similar to these processes. Just as the child learns language "all of a piece," so the holistic reader is engaged in a direct apprehension of the overall adequacy of an essay.

And in a very important sense, a holistic reading really involves a relationship between child and adult that is more than just an analogy. A good holistic reading is a "face to face" encounter, despite the literal absence of one of the participants (White, 1985). One mind (the writer's) is embodied in written expression, and another mind (the reader's) is attempting to see what is being communicated.

A silent dialogue is at work when a person reads through errors that do not impair meaning and concentrates instead on the thought processes at work in a piece of writing. In fact, one could argue that if such a suspension of editorial preoccupation is not accomplished, a substantive reading is impossible. If we are attending primarily to isolated particulars (misspellings, comma mistakes, tense shifts), we will not be able to grasp the overall meaning of an essay. We will instead be engaged in what Polanyi has called a "destructive analysis," which can no more give us an idea of a paper's purpose than a counting of bricks can reveal the shape or function of a building (Polanyi, 1958; Gelwick, 1977)

A half hour's conversation with an experienced holistic reader will indicate the dramatic differences such an "enlightened" reading makes. Such individuals are simply more open to the wide range of strategies that authors can use to attain their purposes. Holistic scoring, by concentrating on organized content rather than such concerns as length, handwriting and sentence level skills, is able to go beyond conventional assessments and

provide a responsible and imaginative framework in which two strangers (writer and reader) are involved in a dialogical process which, at its best, comes close to transcending distance.

In addition, holistic scoring often embodies a developmental awareness that many find quite attractive. At each grade level, emerging ability is given its due. Nor is this developmental sensitivity limited to young and inexperienced writers. Even older and more accomplished composers can profit from a holistic reading of their rough drafts, which are invariably ragged and sloppy - but meaningful - attempts that cry out for mercy and understanding, not for peremptory editorial pruning. Contrary by definition to unrealistic restrictions, holistic scoring helps anyone deeply affected by it to remember that everything finally "perfect" issues from radical imperfection.

Finally, the most recent forms of holistic scoring in large-scale assessment are not "rank-order dependent." That is, they use score point descriptions that are invariant from year to year. A paper earning a 6 in 1990 should earn a 6 in 1993. This invariance in score point descriptions allows evaluators to measure growth despite the inevitable differences in student populations. Some explicators of holistic scoring have pointed to its rank ordering system as a weakness (e.g., White, 1985). That weakness, at least in the large-scale assessment of written expression, no longer exists. This generalizability of score point descriptions also enables readers to become better at their scoring tasks. A dramatic example of reader improvement using the same guide/score point descriptions can be found in Measurement Incorporated's scoring of papers for the New Jersey High School Proficiency Test from 1985 - 1989. 1985 was the first year that MI performed the scoring for the HSPT. Initially, its non-mode-specific 6 point scale created some difficulties even for experienced readers. As a result, room

agreement (in terms of perfect, not adjacent, agreement) was in the upper 50's. Four years later, using the same scoring guide and employing basically the same set of readers, the room average (perfect agreement) was in the low 70's.

It would be hard to overemphasize the importance of this score-scale stability. Able to accurately measure growth in writing across years, educators and administrators are able to identify "problem patterns" that call for changes in curriculum. Ideally, such changes will help make a difference in the ability of students to respond more successfully on a variety of written tasks.

There is no doubt that all of the six reasons I have outlined have contributed substantially to the present popularity of holistic scoring. Nor in my opinion, is it just the case that there is a tension between "is" and "ought" when their reasons are considered. In short, all six reasons are strong ones, and without them any evaluative instrument - not just holistic scoring - would be weakened.

However, despite the popularity of holistic scoring and the good reasons for its success, a "dark side" still exists. This "dark side" consists of a number of different problems which need to be identified and, in many cases, jointly addressed by different groups of concerned professionals. If these problem areas are not dealt with, we may see the return to an exclusive reliance on multiple choice tests, the replacement of holistic instruments by analytic ones, or a combination of indirect and analytic instruments.

The first problem - what "truly" constitutes holistic scoring - is probably the least important of the weaknesses I will discuss, but it is significant enough to warrant consideration, if only to clear the way for the probing of other issues. Actually, this problem is a residue of Cooper's

discussion of holistic scoring in his groundbreaking 1977 study entitled Evaluating Writing. In that work, Cooper argued for an extremely broad definition of holistic scoring that allowed the inclusion of forms of scoring which are mutually exclusive (e.g., analytic and dichotomous scales) to be labeled "holistic." Despite the delineations accomplished in practical terms by means of hundreds of projects in the last 10 years, one still finds references to holistic scoring that are dependent on Cooper's list of holistic instruments. Huot's labeling of three predominant scoring strands (holistic, analytic, and primary trait) is a much more accurate and up-to-date categorization of the practical consensus of the 80's (Huot, 1990). As long as the older labeling remains in effect, fruitless discussions of what is and is not holistic will continue to take place - thus perpetuating confusion and obscuring some very real difficulties.

And, indeed, there are some distressing weaknesses in the practice of holistic scoring. As we turn to a discussion of them, it is important to remember that despite widespread agreement that holistic scoring is a valuable instrument, the enterprise called "holistic scoring" is by no means a unitary phenomenon. When one surveys the assessment scene, the popularity of holistic scoring is obvious, but the general agreement tends to mask some problematic differences. Thus, in a sense, I have come full circle in the writing of this paper. Six years ago, when I first began scoring essays holistically, I was struck by differences. After some familiarity with processes and projects, I started to focus on similarities. Now, once again, differences need to be taken into account - but this time in the service of a higher level of integration in which the widespread agreement can become less discrepant in terms of concrete detail.

One of the most serious weaknesses of holistic scoring may be termed "The Flip Side of Efficiency." That is, while efficiency of scoring is clearly a strength, it can also be a weakness. This weakness can manifest itself in both the testing situation and in the actual scoring of essays. In a number of testing situations, students have only 20 - 30 minutes to write. These tests are relatively easy to administer and do not significantly interrupt the school day.

They do, however, make thoughtful prewriting and significant revision quite difficult. It is remarkable that, given these minimal conditions for composing, students perform as well as they do. But one has to wonder about a number of poor performances on tests of this kind. Is it possible that a significant number of students would do better (in terms of both content and form) if they had more time to think/work through the process of writing? Adequate focus, conceptualization, organization and development often take time. The typical response to this problem is that scoring guides take the restrictions of the test situation into account. Even so, the rather frenzied nature of this form of testing is disturbing, especially when students have been taught the importance of thoughtfulness in the significant "temporal stretch" required for good writing.

Of course, "How much time is enough?" is still a question that admits of no clear answer. We need a number of research studies that focus on the issue and help educators and test makers make more responsible decisions in the large-scale assessment of written expression. It would be especially interesting to compare the results of tests at opposite ends of the spectrum (for example, the ones with a 20 or 30 minute limit and those, like Maryland's state assessment, that allow students an entire day to compose).

The second aspect of the "Flip Side of Efficiency" is that raters can be adversely affected as well. If raters are asked during a 6 or 7 hour workday to read a paper every 2 minutes, they will read from 180 - 200 papers per day. Some individuals pride themselves on being able to process many more papers than this - up to (in incredibly rare situations) 1,000 papers a day. Such assembly-line procedures can have an extremely negative consequence. In the name of scoring efficiency, raters may misscore essays, either because these papers are too unusual and/or original to be dealt with in two minutes or because the sheer monotony of the "forced march" leads to a partial withdrawal of attention.

This problem is compounded by the "lowest bidder" mind set. For example, in an effort to have student writing scored accurately - but cheaply - state departments of education have been known to award contracts to the scoring firm that will cost that state the least amount of money. Naturally, a "bottom dollar" budget means that there is a great deal of pressure on raters to race through packets of papers in the limited number of days allotted to a given project.

Of course, no matter what precautions are taken, some papers are going to be given the wrong score. However, that reality should not be confused with misscoring that occurs because of the sheer insanity of the system. Although holistic scoring has proven to be a reliable system of evaluation, who knows how accurate it could be if its "fast forward" procedures were slowed down somewhat? This is a problem for all professionals concerned with the holistic scoring of writing. One might even say that it is basically an ethical problem that can be put in the form of a rather blunt question: How many individual scores are we willing to sacrifice just so we can "cover the field"

quickly? One other corollary question is "Might it not be better to ensure greater accuracy by being willing to spend a bit more?" These questions have no easy answers, but they need to be posed. Otherwise, holistic scoring will continue to risk being a travesty of the silent dialogue that may be its most outstanding feature.

In terms of interrater reliability and score scale consistency, there is a related problem in situations embodying the distinction between central and local scoring. Central scoring simply means that an outside agency - a testing firm or a scoring center using the constraints enumerated by White - is responsible for the scoring of papers. A classic example of this procedure is a state department of education's decision not to let teachers score the essays written for that state's writing test. To ensure objectivity, the state awards a contract to a firm which agrees to score the papers in a given number of days. If, on the other hand, the state decides to let its teachers score these essays, the scoring is being done locally. Even though finances undoubtedly play a part in the decision for local scoring, many administrators and educators feel that it is important for teachers to have "hands-on" experience in holistic scoring (The Connecticut Mastery Test is a good example).

As one can imagine, the difficulties in replicating the controlled conditions of central scoring in local scoring are considerable. Despite the popularity of holistic scoring in large-scale assessment, quite a few teachers find it foreign to their mode of instruction, which places a greater premium on editorial thinking than holistic scoring does. Even when they are trained to score papers holistically, they have a tendency at some point in the scoring to revert to editorial/analytic preoccupations. However, few states have allocated enough time and money for the identification and retraining (or dismissal) of readers who succumb to these tendencies. In the best central

scoring sites, these quality control practices (checks on the reading in progress) are standard. Daily statistical reports on reader agreement, cross-scoring between different teams of readers, and team-leaders' spot-checking of individual readers are several of the central scoring procedures that are used to monitor "reader drift," that notorious (and usually unconscious) move backward to a reader's own preferences. When that happens, of course, holistic scoring degenerates, and reliability is jeopardized to the core. At the very least, more 2 point discrepancies requiring resolutions are generated. At its worst, this unmonitored drift occurs in the majority of the scoring population, in which case an entire room of readers may be more or less in agreement in giving scores 1 point lower (or higher) than the score point descriptions call for. That is the ultimate nightmare in the large-scale assessment of writing.

Anyone who thinks such a dilemma is contrived should peruse the Auchter and Patience essay already cited (Auchter and Patience, 1989). Since there are dozens of local GED testing sites scattered across the country, these authors are very much aware of the need for - and the difficulty of - replicating the controlled conditions of central scoring. While in my opinion, their solution (numerous recalibration sets guaranteed to cause reader anxiety) is a bit severe, they are clearly right in realizing the seriousness of the situation. Once again, there is no simple answer, but anyone considering local scoring should at least be aware of White's caveats (and suggestions for setting up a holistic scoring session) in Teaching and Assessing Writing (White, 1985). Without some system of judging raters on a consistent basis, the local scoring of papers can come dangerously close to fraud. At the very least, raters should have to qualify on one set of prescored papers, and their work should be spot-checked during live scoring. Ideally, readers would also have to take several recalibration rounds. When

properly done, prescored recalibration sits can be extremely valuable ways of identifying reader drift.

It remains to be seen whether a substantial number of people will begin to pay attention to the problem of central versus local scoring. I suspect that as more states buy into the holistic assessment of writing, the discrepancies in scoring will become more obvious - and more widespread. At any rate, this is an issue that almost by itself could decide the future of holistic scoring in the large-scale assessment of writing. When certain conditions are met, holistic scoring is fairly reliable. When they are not, it is just another mediocre tool. If the bridging of the gap between central and local scoring is not consistently accomplished, holistic scoring will, in too many instances, lose its value as a means of assessing writing - and will probably be replaced by more reliable instruments.

The next weakness of holistic scoring has to do with inappropriate developmental principles. This may seem to be an odd assertion since earlier I praised holistic scoring for its developmental concerns. The contradiction is resolved if one keeps in mind that holistic scoring is not a unitary phenomenon - far from it.

I would like to cite three different examples of this lack of developmental appropriateness. One of these has already been discussed in a global fashion. I am referring to the overemphasis on analytic/editorial concerns. There are several holistic rubrics that in an overzealous attempt to integrate form and content, make impossible demands on youngsters. Just last month I gave a presentation about holistic scoring to a group of New York State teachers and administrators. As we were discussing several fourth grade essays, a number of individuals said that their state rubric did not allow for compositional flaws, and that no matter how organized and developed a

response, it could not earn a 4 (the highest score) if it contained errors in syntax, grammar, or mechanics. This strong perfectionistic streak is, in fact, a part of that rubric (I was given a copy to peruse). It is quite ironic - and unfortunate - that a process that is supposed to emphasize organized content can end up reinforcing the very pedagogical mistake it was meant to correct.

However, one should not suppose that analytic reductionism is the only culprit in this arena. There are other developmentally flawed approaches that commit the opposite error: namely, they glorify organized content and virtually ignore the problem of form. One example of this tendency can be found in the scoring practices New Jersey advocates for its High School Proficiency Test. In many ways, New Jersey's approach is exemplary, since it encourages a non-mode-specific sensitivity to a wide spectrum of student strategies. Students are rewarded for taking compositional risks, many of which involve original combinations of effective strategies.

However, this commendable openness does have a negative aspect. For example, even though the guide says that a paper with a pattern of errors cannot earn a 4 (on a 6 point scale), in actual practice a paper with some organization and development has to be practically indecipherable to receive a 3 (The Registered Holistic Scoring Method for Scoring Student Essays, 1986). Obviously, this form of holistic scoring perpetuates the old form/content dichotomy represented by the classical "red pen" method of evaluation, which emphasizes grammatical correctness. The overemphasis on content merely reverses the importance of the dichotomized items: now content, not form, is apotheosized. The implicit message is that the ability to communicate effectively to a generalized audience (and not just to a group of highly trained raters who can decipher the most garbled of prose) is a rather low

priority item. It seems to me that this overemphasis on organized content is developmentally inappropriate because high school students, in particular, need to begin paying more attention to the conventions of clear communication. After all, whether we like it or not, many employers find rampant misspellings, egregious subject/verb disagreement, and a plethora of run-on sentences quite offensive, and older students - even in a holistic setting - need to be reminded that these things continue to matter in the "real world."

The Riverside Handbook for Focused Holistic Scoring occasionally exemplifies another developmental flaw. This otherwise excellent manual, which consists largely of sample student essays and accompanying comprehensive annotations, too often takes students to task in an unrealistic manner. One of the annotations begins with these sentences: "The writer of this essay (a paper earning a score point of 2 on a 4 point scale) would have benefited from an exposure to the traditional approach to argumentation, especially from a study of logical fallacies. Most of this essay is a categorization of children ... based on oversimplification and a confusion between cause and effect, one of the most basic of logical fallacies." (Handbook for Focused Holistic Scoring, no date given, P. 132). Of course, many adults could also profit from a study of logical fallacies. It seems unfair to apply such a strict standard of logic as correspondence with reality to the writing of students, especially in a large-scale assessment which does not take into account the radical heterogeneity of classroom instruction.

Of course, the three examples just cited should be seen as exemplifying problem areas. I am not singling out any of these tests as such. They are significant only insofar as they indicate conflicting trends within a somewhat misleading overall agreement. If no test were cited, one might think that these very real areas of concern were merely hypothetical situations.

The implications for test makers and state departments of education should be obvious. They should scrutinize the tests they now have (or are planning to implement) and find out what other states/companies are doing. All of this scrutiny and research should be in the increased service of appropriate developmental principles. No test can be perfect, but we have not yet seen the best - that is, the most developmentally sound - holistic measure of writing. I am convinced that, given the success of the flawed, but very good, tools we now have, extremely successful holistic instruments can be devised - and that is one of the challenges we must jointly face during the decade of the 90's. Naturally, progress will be contingent also on selected research that focuses specifically on conceptual and linguistic growth as one moves up the grades and on comparative studies that exhibit developmental strengths and weaknesses of the entire spectrum of large-scale assessments of writing.

The last weakness of holistic scoring - that it has yet to be universally integrated into learning communities - has several components. First of all, the meaning of test scores needs to be explained more fully to educators, students, and parents. The test publishing community has to accept a large part of the responsibility in this area. Far too often, people are simply baffled by the scores that are sent back to them. Even when scores are accompanied by descriptions of score points or the modes of writing tested, very little information about individual papers is given. Why, specifically, did a given paper receive a 4 or a 5? What features of good (or bad) writing did it actually exhibit? Such questions point to the lack of specific diagnostic information provided by a holistic score in the large-scale assessment of writing. Such complaints may not have been voiced as often when the distinction between assessment and instruction was fairly clear cut, but now that the two processes are increasingly seen as closely linked, one hears them everywhere.

Ideally, each student paper holistically scored in large scale assessment would come back with some kind of annotation explaining the score. Given the obvious impossibility of such an enterprise, some publishers have, in additional to individual score reports, offered interpretive manuals which contain a number of annotated student papers. For example, the soon to be published interpretive manuals for the Stanford Writing Assessment Program have annotations at every grade level for all 6 score points in 4 different modes of writing. Such examples, which are accompanied by a detailed description of the modes, a discussion of the expectations for each mode at every grade level, and some annotated samples of analytically scored essays, help bridge the gap between assessment and instruction.

Even so, more has to be done. A first step would be the construction of a holistic "check list," which, if carefully drawn up, could, by providing more detailed information about the holistic score, be able to perform an "annotation function" without taking anywhere near the time that a meaningful annotation would. The cost of scoring services would be increased, but publishers and educators alike need to consider the consequences of not being willing to invest more in a tool that could make a real instructional difference.

Also, holistic scoring has to be incorporated more systematically into the school curriculum. Despite the weaknesses outlined above, holistic scoring represents a vast improvement over the exclusively editorial evaluations still in effect in a large number of classrooms. The NAEP Assessments and others keep hammering away at students' lack of critical thinking skills (Applebee, Langer, and Millis, 1986). With its emphasis on organized content, holistic scoring is the perfect vehicle in the facilitation of higher-order thinking (Braungart-Bloom, 1986). And yet, although holistic scoring has shown itself

to be a successful measure of writing, too many administrators and teachers refuse to give it a chance, even in English/Language Arts classrooms. I would like to see educators everywhere examine the holistic writing tests that are available, pick one - or devise one themselves - and find ways to implement the holistic scoring of writing in concrete classroom situations.

This means all classrooms and all grade levels. There is no reason for holistic scoring to become the province of English and Language Arts teachers. Educators in other disciplines should recognize the freeing-up effect of holistic scoring. The evaluation of writing no longer needs to be "English Department business" only, because with some responsible training, competent teachers can evaluate the kinds of organization and content development required in their particular disciplines.

It needs to be emphasized also that teacher empowerment is not the only benefit of such integration. Students can be taught the 4 or 6 point scale and can learn in very concrete terms how to evaluate their essays and those of their peers. I think, too, that holistic rubrics and assignments could, as the year goes along, change at every grade level to reflect the kinds of audience-oriented conventions of language that are appropriate to given levels of maturity. In short, higher and higher levels of integration only dreamed of can be attained if all of us - publishers, researchers and educators - take a hard look at what we have and commit ourselves to the fine-tuning of an already effective measurement instrument.

The preceding discussion prepares the way nicely for the enhancement of portfolio assessment. Holistic scoring can be criticized for its "snapshot" nature; one only sees a sample or two of any given student's writing. Holistic portfolio assessment would help place writing squarely in the context of the meaningful "temporal stretch" from September to June and emphasize that

written expression is an ongoing organic enterprise - also that recursiveness is a critical part of this whole. Quite often, what a student takes to be the end of a piece opens out into a new beginning.

In terms of meaningful integration, one more issue has to be mentioned. It is not peculiar to holistic scoring, but the fate of holistic scoring may well be bound up with it. In the respective communities of education and publishing, especially, one often finds a parochial protection of concerns that is an impediment to progress. For example, as an employee of a private testing firm, I had to sign a disclaimer for each project in which I promised not to give away any secrets that had to remain secure. But exactly what kinds of secrets would they be? And if numerous yet-to-be specified doors stay closed, how can researchers, educators, and publishers work together? We are in desperate need of interdisciplinary dialogue, which alone can enable us to make a recognizable difference in students' ability to express themselves on paper. Obviously, if information and ideas cannot be shared without some ominous background threat of legal intervention, competition and mistrust will never stop being the order of the day. Everyone interested in the improvement of both assessment and instruction should be prepared to address this critical issue of test security. What, if anything, can be done to respect certain needs for security of material without continuing to perpetuate the counterproductive isolationism that tends to characterize the present situation?

As I have shown, holistic scoring has its weaknesses, but I am convinced that, given greater conditions of openness, it can become a very impressive assessment and instructional instrument. If I had to give it a score, I would say it has earned so far a mid-to-upper-range 4 on a 6 point scale. It's up to us - all of us - as its joint authors, to compose a better organized and more richly detailed text that can instruct and satisfy generations to come.

Reference List

Applebee, A.N., Longer, J.A., and Mullis, I.V.S. (1986). The writing report card: Writing achievement in american Schools (ETS Report No. 15-W-02). Princeton, NJ: Educational Testing Service.

Auchter, J.C., and Patience, W. (1989, March). Decentralized large scale essay scoring: Methods for establishing and evaluating score scale stability and reading reliability. Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco.

Barrett, W. (1978). The illusion of technique: A search for meaning in a technological civilization. New York: Doubleday.

Braungart-Bloom, D.S. (1986, April). Assessing higher thinking skills through writing. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Breland, H., & Jones, R.J. (1984). Perceptions of writing skills. Written Communication, I, 101-109.

Bronowski, J. (1965). Science and human values. New York: Harper and Row.

Cooper, C.R., & Odell, L. (1977). Evaluating writing: Describing, measuring judging. Urbana, IL: National Council of Teachers of English.

Diederich, P.B. (1974). Measuring growth in English. Urbana, IL: National Council of Teachers of English.

Faigley, L., Cherry, R.D., Joliffe, D.A., & Skinner, A.M. (1985). Assessing writers' knowledge and processes of composing. Norwood, NJ: Ablex.

Farr, M. (1984). Writing growth in young children: What we are learning from research. In C.J. Thaiss and C. Suhor (Eds.), Speaking and Writing, K-12 (pp. 126-143). Urbana, IL: The National Council of Teachers of English.

Freedman, S.W. (1979). How characteristics of student essays influence teachers' evaluation. Journal of Educational Psychology, 71, 328-338.

Gelwick, R. (1977). The way of discovery: An introduction to the thought of Michael Polanyi.

Hake, R. (1986). How do we judge what they write? In K.L. Greenberg, H.S. Wiener, & R.A. Donovan (Eds.), Writing assessment: Issues and strategies (pp. 153-167). New York: Longman.

Handbook for focused holistic scoring (no date or place of publication given). The Riverside Publishing Company.

Hartse, J.C., Burke, C.L., & Woodward V. (1982). Children, their language and world: Initial encounters With print. NIE final report (NIE-G-79-0132).

Hogan, T.P., and Mishler, C. (no date given) Relationships among measures of writing skills. A review prepared for the National Assessment of Educational Progress.

Huot, B. (1990). The literature of direct writing assessment: Major conerns and prevailing trends. Review of Educational Research, 60, 237-263.

Kaufmann, W. (1958). Critique of religion and philosophy. New York: Doubleday.

Matson, F.W. (1966). The broken image: Man, science and society. New York: Doubleday.

Polanyi, M. (1958). Personal Knowledge; Towards a postcritical philosophy. Chicago: University of Chicago Press.

The registered holistic scoring method for scoring student essays (1986). N.J.: New Jersey State Department of Education.

Stanford Writing Assessment Program, Second Edition: National Norms Booklet. (1990) San Antonio, TX: The Psychological Corporation.

Stock, P.L. & Robinson, J.L. (1987). Taking on testing: Teachers as testers and researchers.

White, E.M. (1985). Teaching and assessing writing. San Francisco: Jossey-Bass.

Wild, G. (1970). The Radical Empiricism of William James. New York: Doubleday.

Wiser, B., Kelleher, A., and Rolfe, D. (1989, February). Contrasting developmental scales for four types of writing. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.